Forbidden Synonymous Substitutions in Coding Regions[1]

Roy J. Britten

Division of Biology, California Institute of Technology, and Department of Ecology and Evolutionary Biology, University of California, Irvine

In the evolution of highly conserved genes, a few "synonymous" substitutions at third bases that would not alter the protein sequence are forbidden or very rare, presumably as a result of functional requirements of the gene or the messenger RNA. Another 10% or 20% of codons are significantly less variable by synonymous substitution than are the majority of codons. The changes that occur at the majority of third bases are subject to codon usage restrictions. These usage restrictions control sequence similarities between very distant genes. For example, 70% of third bases are identical in calmodulin genes of man and trypanosome. Third-base similarities of distant genes for conserved proteins are mathematically predicted, on the basis of the G+C composition of third bases. These observations indicate the need for reexamination of methods used to calculate synonymous substitutions.

Introduction

As a result of the degeneracy of the genetic code, a base substitution in a coding region may yield a different codon that specifies an unchanged amino acid, now termed "synonymous," or "silent," substitutions. Most of these changes occur at third bases in codons. Because they do not affect the amino acid sequence, many such base changes have been incorporated with relatively high frequency during evolution of gene coding regions. The resulting sequence drift is very useful in assessing interspecies relationships and in studying evolution and systematics (e.g., see Li and Graur 1991, pp. 67–131). It is well recognized that there are restrictions on possible substitutions at these positions. For example, different codons in a degenerate set are not used with equal frequency, a situation termed "codon bias" (e.g., see Grantham et al. 1980). A part of the bias probably results from selection for translation efficiency of highly expressed genes (Ikemura 1982), but the reasons for most bias are not yet understood. The selections causing bias might occur on various levels, even indirectly through the mechanisms of mutation (Wolfe et al. 1989). The patterns of codon usage and the resulting G+C content of third bases vary between classes of organisms, between related species, between regions of individual genomes (Bernardi and Bernardi 1985), between individual genes (Sharp et al. 1989), and even within a given gene (Lawrence et al. 1991). There are additional and sometimes severe restrictions on synonymous substitutions at a few particular positions in a coding sequence, restrictions that are not due to recognizable codon bias (Bains 1987), and new aspects of this phenomenon, including the identification of large classes of slowly changing third bases, are examined in the present paper. These slowly changing third bases may be significant either to

1. Key words: rare synonymous substitutions, mutation rate, actin gene evolution, codon usage, third-base G+C composition.

Address for correspondence and reprints: Roy J. Britten, Kerckhoff Marine Laboratory, 101 Dahlia Avenue, Corona del Mar, California 92625.

Mol. Biol. Evol. 10(1):205–220. 1993.

0737-4038/93/1001-0012$02.00

messenger RNA structure or to requirements of the coding sequences (e.g., see Lawrence et al. 1991). Specific slowly changing third bases also affect the observed rates of drift of DNA sequences and may explain part of the large apparent drift-rate contrasts observed among genes and species (e.g., see Britten 1986).

The quantitative determination of the rates of synonymous substitution (Li et al. 1985) is important for interspecies comparisons of genes, and corrections are often made for expected numbers of multiple substitutions at the same positions. These are significantly affected by restrictions on possible substitutions (Sharp and Li 1987; Bulmer 1991; Bulmer et al. 1991; Long and Gillespie 1991). The present paper reports that the percent identity at third bases is inversely correlated with the difference in G+C composition at third bases. A mathematical relationship is established between the percent identity at third bases for distant genes and the third-base G+C composition of the compared genes. The average difference in G+C composition at third bases in 1,275 comparisons is ~25% and varies widely even among closely related actin genes. This is partly due to interspecies differences in codon bias and partly due to gene or regional differences in composition. The detection of forbidden or rare substitutions at specific third-base positions is influenced by other restrictions on third-base usage. Thus, the overall synonymous-substitution divergence and the effects of G+C composition are described first.

Methods

The sets of sequences for comparison were obtained by searching GenBank with a complete actin DNA sequence, using FASTA (Pearson and Lipman 1988). A FORTRAN program was written that extracted the sets of sequences from the FASTA alignments and formatted them. Pseudogenes and duplicates and partial sequences were manually removed. The set was then aligned using CLUSTALV (Higgins and Sharp 1988) on a Sparc-1 Sun workstation. These alignments were manually corrected to remove inserts that interrupted codons. A FORTRAN program was written that read these alignments, translated the sequences, made all possible comparisons, and calculated all of the needed parameters such as GC3% (percent G+C content in the third codon position), etc.

Results

Divergence of Third Bases, Their G+C Composition, and Codon Usage

Figure 1 is a graph of the similarity at third bases, as a function of the AAD% (percent amino acid sequence difference) for ~51 actin genes from eukaryotes. Most of the pattern is the end result of long periods of drift, since most of these actins are very distant from each other in terms of synonymous substitutions. The vertical axis is the percent identity at third bases and the horizontal is the AAD% for 1,275 gene comparisons. The number of comparisons that fall into each box (2% vertically and 1% horizontally) is printed. Ideally, the synonymous-substitution differences among these genes should be compared with the time of evolutionary separation of the lineages of the genes, but, in general, this information is not available or is relatively crude. However, the AAD% is usefully related to the time since the separation of the gene lineages, though it is affected by selection against amino acid replacements, and this may vary from case to case. Figure 1 does have the advantage that it depends only on the direct sequence comparison rather than on any models or indirect approaches.

The most obvious feature of figure 1 is the large spread of synonymous-substitution similarities. Little of this vertical spread is a result of statistical fluctuations. Actin

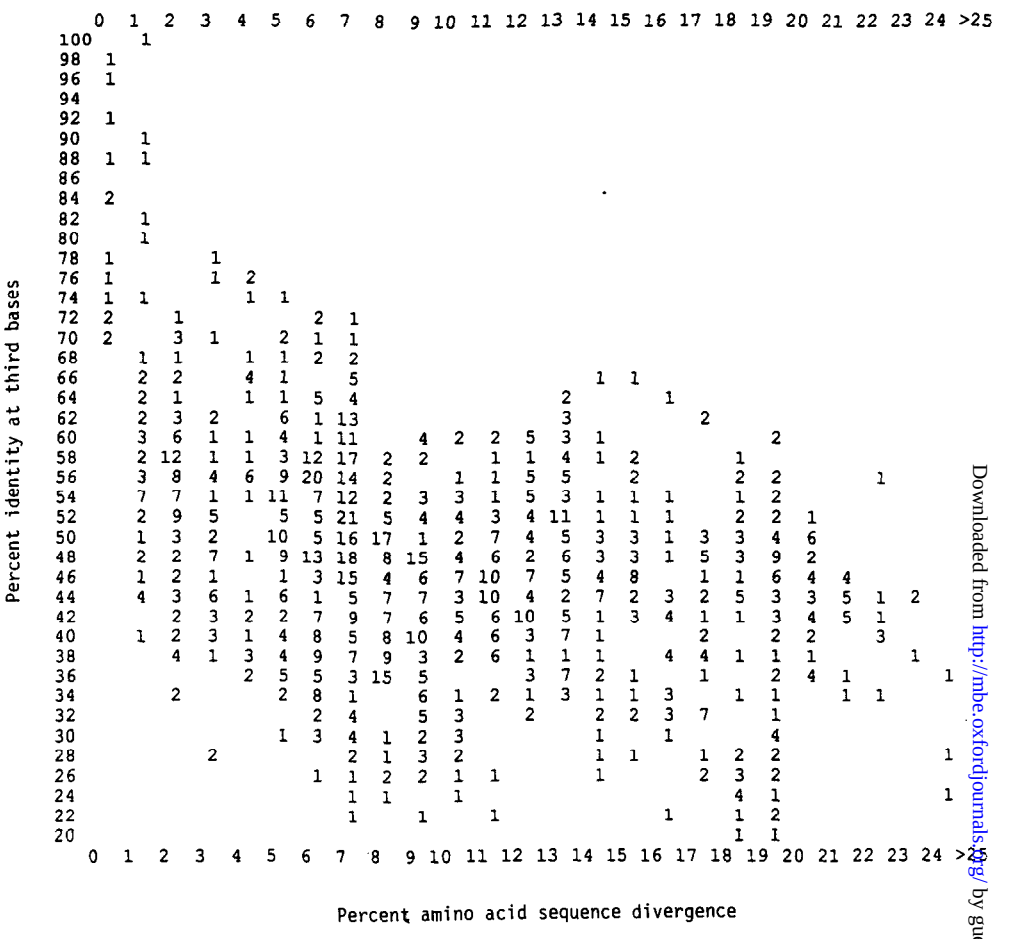

FIG. 1.—Percent identity at third bases, for coding regions of actin genes, as a function of amino acid divergence of the proteins. All pairs of 51 actin genes are compared for a set of 1,275 comparisons. The abscissa is the percent amino acid divergence for the implied protein. Boxes are formed by dividing the abscissa into 1% intervals and the ordinate into 2% intervals. The number of comparisons falling into each box is listed. The sequences were drawn from GenBank by using FASTA (Pearson and Lipman 1988) and searching with a typical coding region. Special FORTRAN programs removed duplicates and reformatted, while pseudogenes were removed by hand.

includes 375 codons, so points in the central part of figure 1 each represent 150–200 third-base substitution differences, and the statistically expected vertical spread is ~7%. The much larger observed spread is correlated with the base composition of the genes being compared, which in turn is due to the codon bias patterns of the individual genes. For the purposes of this paper, important aspects of the codon usage patterns can be represented simply by the average percent G+C composition at third bases (expressed as GC3%). The simplifications that result permit a picture of the principal relationships.

Figure 2 shows the strong effect of the difference in G+C composition on the percent of third-base identity between the compared genes. Genes with similar but extreme composition in the third base show a reduced amount of synonymous-substitution divergence, while genes with very different composition typically show large

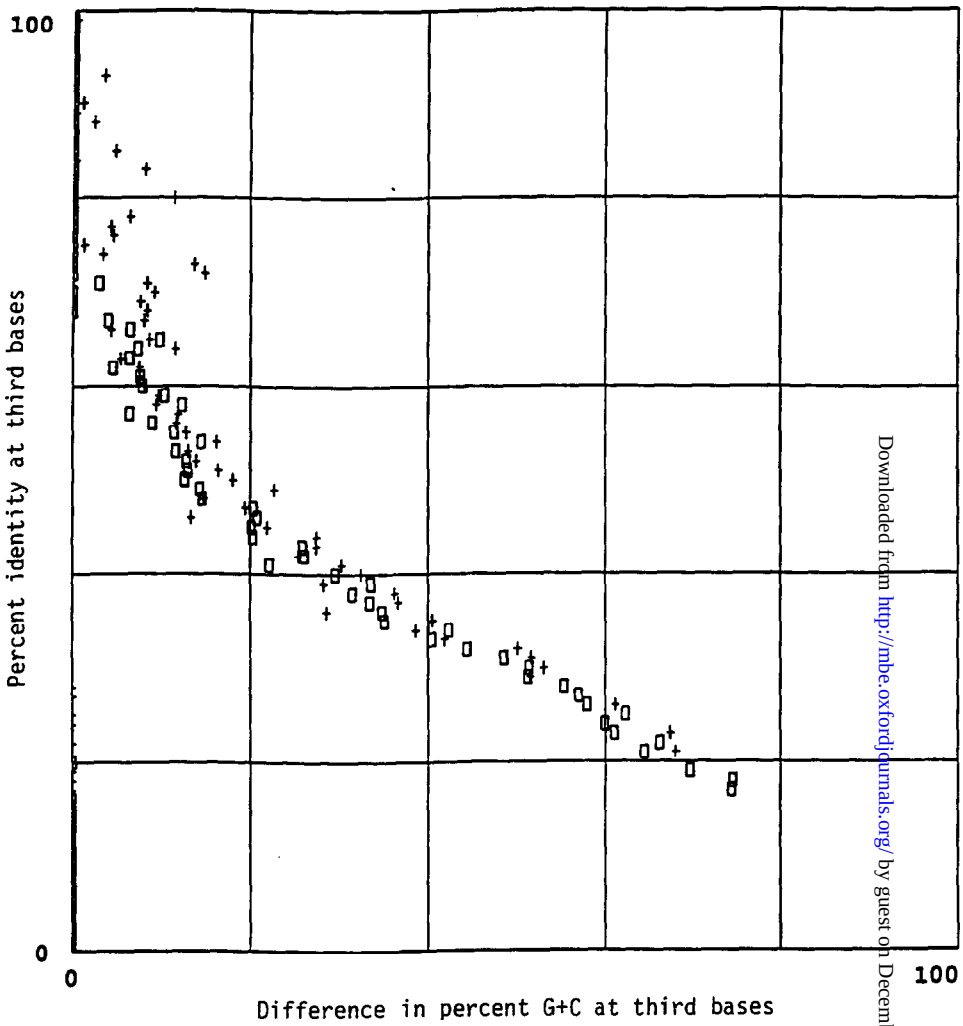

FIG. 2.—Third-base identity as a function of difference in third-base composition for 1,275 actin gene comparisons. For each comparison shown in fig. 1, the difference in percent G+C at the third base (i.e., GC3%) was averaged for comparisons having the same percent identity at third bases (±0.5%). The abscissa is the average of the differences in third-base composition (which have a wide spread). Boxes denote comparisons with amino acid sequence divergence >7%, and plus signs denote that <7%. In the upper region are comparisons for distant actins (boxes) that have similar but extremely high (or low) G+C in their third bases. They have many identical third bases, owing to restricted usage, as for the calmodulin case in table 1. These data show a correlation between average difference in composition and divergence of third bases. Large differences in composition are very common. For example, the average difference in third-base composition (i.e., GC3%) is 30% for comparisons with 40% third-base identity. This example was chosen because 40% third-base identity is the expected steady-state value at great distance for typical actin codon usage. Obviously, few fit a simple model of drift with usage restriction that underlies many synonymous substitution calculations.

amounts of synonymous-substitution divergence. Particularly interesting are the upper points (boxes) for genes with large (7%–25%) AAD% which nevertheless show great similarity (60%–70%) at third bases. Actin genes are very conserved, and mammalian muscle and cytoskeletal actin amino acid sequences differ by <7%. The observed 60%–

70% identity of third bases is for very distant genes whose last common ancestors may have existed in the Precambrian. Most of these cases are comparisons of genes that have very similar base composition and very high GC3% at third bases; a few are cases of very low GC3% at third bases in both genes. The cause of the limited synonymous substitution appears to be the restricted codon usage that is required to maintain very high or very low G+C composition. That can be considered to be convergence or simply the limited frequency of the bases in each position, maintained by a very restricted set of forward mutations and balanced by reverse mutations. For perspective it is worth noting that genes with similar but medium values of G+C composition of third bases (say, near 40%) are expected to have a medium percent identity at third bases, corresponding to the classical steady value resulting from a balance between forward and reverse mutations, depending on the codon bias.

The synonymous similarity at great distances is traditionally expected to be the net result of a balance between mutations that cause differences and their reverse. For example, in a model Monte Carlo calculation the DNA sequence of an actin gene was substituted at rates of individual codon substitution that were selected to maintain an average usage pattern (with amino acid replacements forbidden). The steady-state value was 40% identity of third bases after long periods of divergence, right in the middle of the distribution in figure 1. However, figure 2 indicates that this is not the appropriate model for typical actin gene evolution, because, at 40% identity of third bases, the average difference of third-base G+C composition is ~30%. For individual comparisons there is a range of 2%–50% difference in G+C composition for cases with ~40% identity of third bases (not shown). These observations have an impact on attempts to assay the amount of divergence at third bases in order to assess evolutionary relationships between species. It also shows that G+C composition difference between compared genes is typically very large, and this observation has implications for genomic evolution (see Shields et al. 1988).

The effect of the sharing of high GC3% is shown by the calmodulin example mentioned in the abstract and summarized in table 1. Like actin, calmodulin is a highly conserved gene, and the alignments are excellent. The large amino acid divergences (for the highly conserved calmodulin) rule out suggestions of interspecies transfer of a gene even though *Trypanosoma cruzi* is a human parasite. What these genes share is high GC3%, which must be the cause of the high percent identity of third bases. In table 2 are the compositions and amino acid sequence divergence for the genes of the six human actin types. Both cytoplasmic actins have high GC3%, while one of the

Table 1
Excess Synonymous Similarity for Calmodulins That Is Due to Codon Usage Pattern

Genes Compared (GC3%)[a]	Synonymous Similarity[b]	AAD%
Human high (91) vs. trypanosome (91)	70%	19
Human low (31) vs. trypanosome (91)	31%	10
Human high (91) vs. human low (31)	29%	15

[a] "Trypanosome" designates GenBank sequence TRCCALB2, "human high" designates GenBank sequence HUMNB1, which has 91 GC3%, and "human low" designates HUMCAM with 31 GC3%.

[b] Percent identity at third bases for matching amino acids.

Table 2
Composition and Relationships of Genes for Six Human Actin Types

Actin Type	GC3%	CpG3%	AAD%[a]
Cytoplasmic beta	84.1	13.3	1.3
Cytoskeletal gamma	85.7	12.5	
Skeletal muscle alpha	89.1	17.8	1.1
Alpha-cardiac	66.6	5.1	
Smooth-muscle alpha	64.2	4.0	0.8
Smooth-muscle gamma	69.2	2.7	

[a] The AAD% between the two cytoplasmic actins and all of the muscle actins is 6%, and that between the two smooth-muscle actins and the other two muscle actins is 2%.

muscle actins has high GC3%. The codon usage pattern is very different for two closely related muscle actins.

Another example of the effect of high GC3% is the comparison between two human actins that have high and low GC3% and an actin from the mollusk *Aplysia californica* that has high GC3%. All three of these actins diverge from each other, in amino acid sequence, by only slightly >6%. Human skeletal muscle actin (89 GC3%) has a 69% third-base identity with the *Aplysia* actin (80 GC3%), while human vascular-smooth-muscle actin (64 GC3%) shows only 55% third-base identity with the *Aplysia* actin. It is clear that the examples with high GC3% show high percent identity of third bases even though they are very distant genes. Table 2 shows a large difference in GC3% for the two fairly closely related actins (skeletal 89 GC3% and cardiac 67 GC3%), which have only a 1.06% AAD% between them. This is an example of the typical large GC3% difference that is not due to the difference in codon usage between species, since they are both human genes and since a similar pair of genes are present in mouse. Whatever the underlying mechanism of changes in G+C composition, the data in the tables and the correlation in figure 2 indicate that third-base composition and codon usage patterns are the main reasons for the vertical distribution of points in figure 1. An equation representing these relationships will be described after a brief discussion of CpG selection.

Selection for CpGs

Gruskin et al. (1987) observed a large number of CpGs held in common between an eel calmodulin gene and an intronless chicken calmodulin gene. Their suggestion that the CpGs may be shared as a result of a virus-carried horizontal transfer of a gene actually initiated these studies of the correlations between both GC3% and CpG3% (percentage of third bases that are C's that are part of CpGs) and synonymous-substitution differences. For this reason table 2 includes the CpG3%. There is evidence suggesting that the CpGs are a necessary concomitant of the high GC3% and that this is the probable reason for CpGs being held in common between distant genes. Generally, among vertebrates the CpG% falls far below the random expectation for dinucleotides, apparently because of methylation and high C-to-T mutation rate in CpG among vertebrates (Coulendre et al. 1978). However, when codon usage leads to high GC3%, the case differs, and in such codons the C's present in third-base positions often are automatically parts of CpGs, as a result of the amino acid sequence. Strong selective forces are required to maintain the high GC3% and CpG3%. There

may be reduced local methylation (Cooper and Krawczak 1989; Boyes and Bird 1992), a possibility requiring the selection of complex chromosomal mechanisms in local regions.

Mathematical Representation

The following equation describes the observed amount of synonymous similarity, based only on the third-base percent G+C3% of the two compared genes:

$$S = a(1-v+w)[1-2(1-v)w]+c. \quad (1)$$

Here S is the fraction of third bases that are identical, and v and w are the third-base composition (fraction G+C) of the two compared genes, with v being the higher of the two. The numerical constants (a = 0.95 and c = −0.07) are small corrections chosen for the best fit to the data of figure 1. The good agreement between observed and calculated third-base sequence identity is shown in figure 3. The two terms of the equation (included in parentheses and brackets) have direct significance. The first $(1-v+w)$ simply goes to zero when the two compositions have the maximum possible difference and linearly rises to 1 as the two compositions approach each other, incorporating the observation shown in figure 2. The second term $[1-2(1-v)w]$ sets a minimum amount of divergence (due to the restricted codon usage) when the compositions are either large together or small together. This term is 1 when $v = w = 0$ or $v = w = 1$ and is just 0.5 when $v = w = 0.5$. Multiplied together, these terms describe the observed effect of base composition over the whole range of data in figure 1.

Equation (1) fits the data well, with an average deviation of 0.043 for >1,000 comparisons. The comparisons shown in figure 3 are for actin and have been repeated for calmodulin, histone H3, and ubiquitin, with similar results, although the constants a and c vary slightly. The constants evaluated for actin, shown in equation (1), are preferable for possible general use, since actin contains more codons and the accuracy is better. The direct effects of codon usage expressed in equation (1) explain the wide spread of synonymous-substitution differences.

A few individual gene comparisons do not follow the general pattern, perhaps in part because of codon usage restrictions not reflected in the G+C composition of the third bases. The most striking outliers include comparison of yeast actins showing 71% third-base identity and ~6 AAD%, but without GC3% quite low enough to explain the similarity. A similar observation is true for yeast ubiquitins. A conserved fragment of Hox genes was compared, but it is short and the GC3% is consistently high, so it is not suitable as a test of equation (1). There is further evidence that other phenomena influence the differences between synonymous positions in compared genes. In data not presented, Ca^{++} ATPase genes show unexpectedly high third-base identity among mammals that is apparently not due to high or low G+C composition. These outliers, for which other processes may be significant, are infrequent, although they do form a risk in the use of third bases to assess interspecies relationships. The other highly conserved genes that I have examined (histone H3, calmodulin, and ubiquitin) show patterns that are similar to those seen in figures 1 and 2, and, in particular, the strong correlation between the difference in third-base composition and amount of synonymous-substitution divergence is observed in each set of genes. For these genes, as mentioned, equation (1) fits the data well.

Figure 4 shows the result of using equation (1) to reduce the effect of G+C composition of third bases, for the data of figure 1. The average of the observed minus

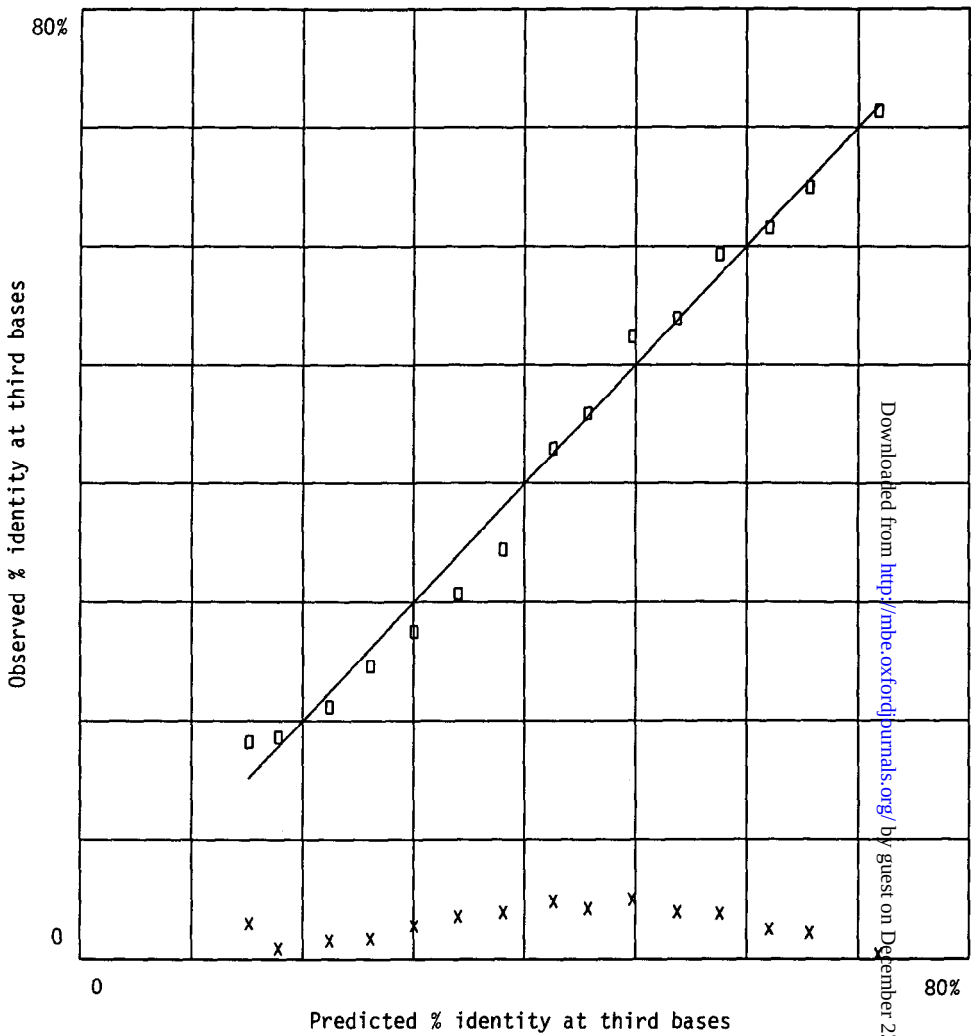

FIG. 3.—Predicted vs. observed percent identity at third bases, using eq. (1). For comparisons shown in fig. 1, equation (1) was used to calculate the expected third-base similarity on the basis of the GC3% of the compared genes. The observed and predicted values (boxes) were averaged for each interval on the abscissa, and the mean deviation (×) was calculated. All comparisons with <3% amino acid divergence were eliminated. Of all 1,275 values, there are a very few outliers that may show a statistically significant deviation between observed and predicted values.

the predicted third-base percent identity is plotted against the amino acid divergence of actin genes. The statistical scatter is reduced by the averaging, and it becomes clear that the actin genes do not follow the classical expectation that the percent third-base identity falls exponentially to the steady-state value controlled by codon bias. A slowly changing component appears that corresponds to 10%–20% of the total positions. Since these slowly changing positions are a minority, it seems likely that slow rate of change is due to specific restrictions on the change of these bases, rather than being a result of overall codon bias; these restrictions are examined by a different approach in the following section.

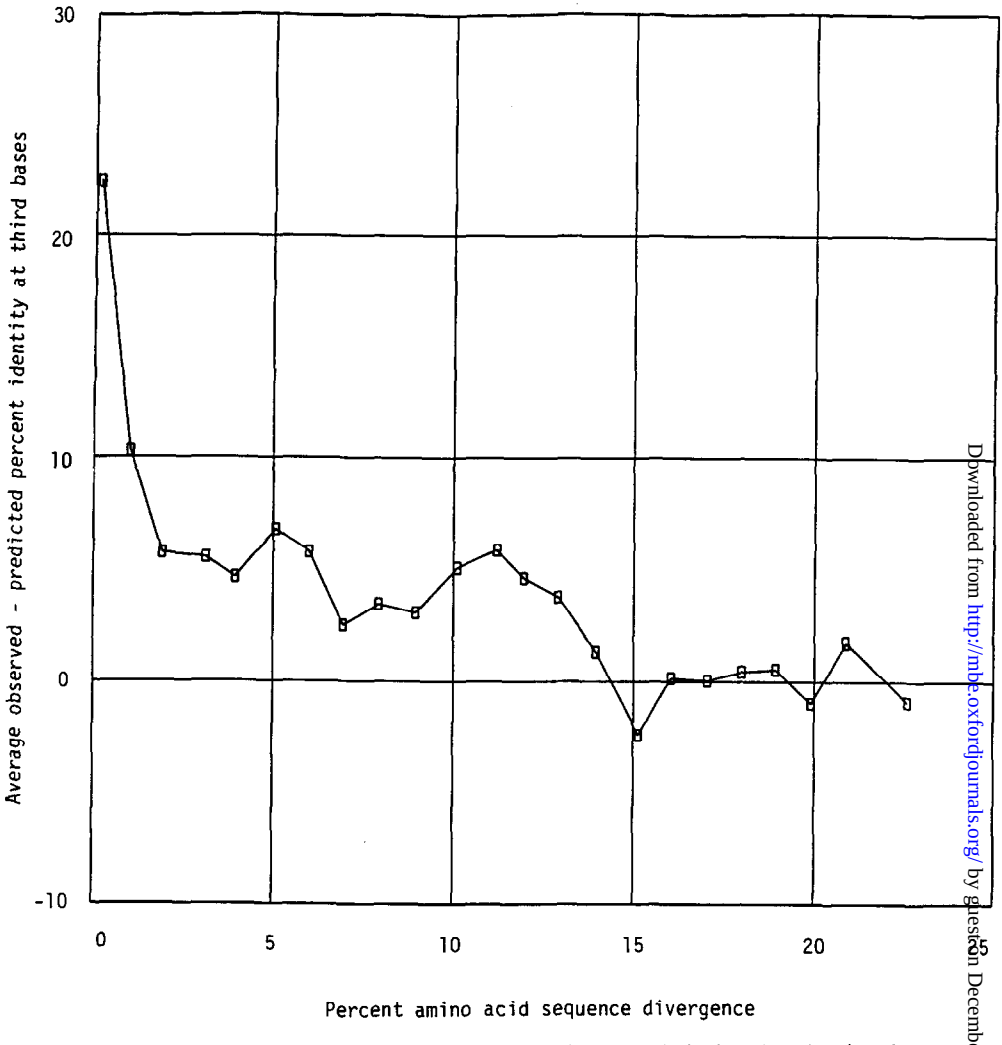

FIG. 4.—Average of observed minus predicted percent third-base similarity, plotted against the percent amino acid sequence divergence (i.e., AAD%) for a set of actin gene comparisons. A set of actin genes have been compared as in fig. 1, except that the yeast genes have not been included. Comparisons were collected in sets for intervals of AAD%—from 0%–0.5%, 0.5%–1.5%, etc.—and the AAD% was averaged for the abscissa. For each comparison the predicted percent third-base similarity for distant genes was calculated using eq. (1) (with $c = -0.095$). This predicted value was subtracted from the observed value and then was averaged for each set. The average observed values are, as expected, higher than the predicted values for closely related genes, because limited extents of substitution have occurred. However, the difference does not exponentially fall to zero. Instead, after a rapid initial drop, there is a slow fall, reflecting the effect of the slowly changing third bases exhibited in fig. 5. This phenomenon affects standard corrections for reversion in synonymous-substitution comparisons.

The Forbidden or Rare Synonymous Substitutions

The purpose is to estimate the probability of events of change in third bases (for a conserved set of genes that are distant from each other), in order to identify positions that change slowly or not at all. The first step was to align a set of 39 actin genes (table 3) and determine the majority third base for each codon. Then the bases that differ

Table 3
List of Actin Genes Used for Fig. 5

Common Name of Group	S_s[a] (%)	AAD%[b]	GenBank Identifier	Description
Human	0.0	0.0	HUMACTASK	Human skeletal muscle alpha
Rodent	71.4	0.0	MUSACSM	Mouse skeletal muscle
Bird	71.4	7.4	GOOACTB	Goose beta-actin
Human	68.0	6.9	HUMACTCGR	Human cytoskeletal gamma
Mollusk	68.8	6.6	APLACTIN	*Aplysia californica* actin
Human	57.7	2.4	HUMSMGA	Human enteric smooth muscle
Rodent	55.4	2.1	MUSACTASM	Mouse vascular smooth muscle
Rodent	54.9	2.1	RATACTAV	Rat vascular alpha
Rodent	52.1	1.1	MUSACTCM	Mouse alpha-cardiac
Toad	54.2	1.3	XELACACR	*Xenopus laevis* alpha-cardiac
Human	54.9	2.1	HUMACTA	Human vascular smooth muscle
Dipteran	62.3	6.9	DROACT87E	*Drosophila melanogaster* actin gene
Rodent	61.0	7.1	MUSACTBR	Mouse cytoskel beta-actin
Rodent	54.3	2.4	M26689	Mouse actin
Dipteran	64.2	7.7	DROACT8F	*D. melanogaster* actin gene
Silk moth	64.7	7.7	BMOACTA1	*Bombyx mori* gene actin 1
Rodent	60.8	7.4	MUSACTMEL	Mouse A-X actin
Rodent	53.7	2.4	RATACTGE	Rat gamma-enteric smooth muscle
Bird	59.3	7.4	CHKACCY5	Chicken type-5 cytoplasmic
Rodent	58.1	6.9	RATGAMACT	Rat cytoplasmic-gamma
Rodent	58.4	6.9	MUSACTGCS	Mouse cytoskeletal gamma-actin
Toad	47.3	0.8	XELACTA22	*X. laevis* sarcomeric alpha
Silk moth	59.7	7.7	BMOACTA2	*B. mori* gene actin 2
Dipteran	57.7	7.2	DROACT2A	*D. melanogaster* actin gene
Toad	44.8	1.1	XELACTA2	*X. tropicalis* sarcomeric alpha
Dipteran	60.1	7.2	DROACT5CX	*D. melanogaster* actin gene
Toad	43.5	1.1	XELACASR	*X. laevis* alpha skeletal
Toad	47.1	7.1	M24769	Type-5 *X. laevis*
Oomycetes	58.8	15.4	PHTACTA	*P. infestans* actin (actA) gene
Toad	46.5	7.1	M24770	Type-5 *X. laevis*
Plant	51.6	12.5	RICRAC1	*Oryza sativa* RAc1 actin
Tunicate	36.5	4.5	SCLMUSACT	*S. clava* muscle actin
Shrimp	40.0	5.9	SHRACT403	*Artemia* actin
Tunicate	41.1	5.6	SCLMUSACU	*S. plicata* muscle actin
Shrimp	37.3	7.4	SHRACT211	*Artemia* actin
Zygomycetes	42.7	11.6	ABGACT2	*Absidia glauca* actin (ACT2)
Oomycetes	51.8	20.4	PHTACT	*Phytophthora megasperma* actin
Oomycetes	46.8	21.4	PHTACTB	*P. infestans* actin (actB) gene
Shrimp	35.2	7.4	SHRACT205	*Artemia* actin

[a] Percent identity of third bases to human skeletal muscle actin.
[b] Versus human skeletal muscle actin.

from the majority were identified. The next step was to add the number of such differing bases for each position in the actin-coding sequence. Those positions that have low sums are identified as the specific positions that are slowly varying. Three positions do not vary at all; that is, all actin genes in the set of 39 have the same base at the third position. These are third bases (all C) in codons for one ASN and two PHEs. Among this set of 39 actin genes, the usage of C in the third base is 67% for ASN and 77% for PHE. Thus the simplest estimates of chance occurrence are 0.67^{39}

and 0.77[39], or $\sim 10^{-7}$ and $\sim 10^{-5}$, respectively, per position. Thus the probability of the random occurrence of three such nonvarying positions in this set is very small (10^{-17}), even when restricted usage is taken into account. There are also five positions that show only one variant in this set of 39 actin genes. The probability of chance occurrence is also very low for this set. There are 57 positions that have fewer than eight variants, and most of these are probably significant, although some may be due to chance variation. This will be further examined below.

In figure 5, the nonvarying or slightly changing positions are shown at their positions in the actin sequence. In this figure, certain slightly changing positions are clustered together. For example, at the 3′ end of the gene is a cluster of four positions that includes one position that has one variant in the whole set of 39, two that have two variants, and one that has four to seven variants. Eight codons 5′-ward is a cluster of seven adjacent positions, including one unchanging position and six of the 4–7 and 8–11 classes. It is fairly probable that members of the 8–11 class occur by chance, and Monte Carlo models suggest that perhaps ⅓–½ of the members of the 8–11 class are due to background chance events. However, there is no reason that these background cases would cluster together, and thus clustered members of the 8–11 class are probably significant.

Because of the complexity of a statistical calculation, a Monte Carlo computer model was chosen. To match the data of figure 5, 39 copies of the reference sequence were mutated at random to different extents, rejecting all mutations that changed amino acids. The distribution of mutations was chosen to give the observed number of variants at twofold-degenerate positions. In the resulting set of data, no third-base positions had fewer than nine variants. Thus no positions fell in the classes for the upper four lines in figure 5, and there is little doubt that the 57 positions in the upper four classes are the result of locally specific restrictions on the amount of variation of these positions. If this is so, ~15% of the third bases of codons in actin genes are specifically restricted in variance. This result is consistent with the number implied by figure 4. The specific restrictions on change at these positions are in addition to the codon usage restrictions that are effective at most positions.

A study similar to that shown in figure 5 has been made with a set of 55 histone H3 genes. It showed that there are 15 third-base positions that have a reduced degree of variation and ~5 positions that can be classified as having very rare variation, i.e., one or two variations among the 55 genes examined. None of the third-base positions were absolutely conserved. Studies with 43 ubiquitin genes gave a somewhat different result. There are ~10 positions that show reduced variation, but none (Bains 1987) for which the variation is rare or absent. There are five positions that do not vary among 13 close relatives of the reference sequence (a human ubiquitin). These third bases vary among a large number of other ubiquitins, implying that, if there are functional roles for ubiquitin third bases, then these roles have changed during ubiquitin evolution and may be fulfilled by other bases. The fact that the ubiquitin gene occurs in clusters of linked sequences probably is a part of the reason for its difference from the other genes examined. Studies of 31 calmodulin genes show that there are two positions that have very rare variation. For one third-base position, only one gene differs from the majority; for the other position, two genes differ from the majority. However, there are 18 positions that show variation that is significantly less than that observed for the majority of positions. In summary, of the four genes examined, all appear to contain 10%–20% of positions for which the third-base variation is reduced because of specific requirements at these positions, as opposed to general usage re-

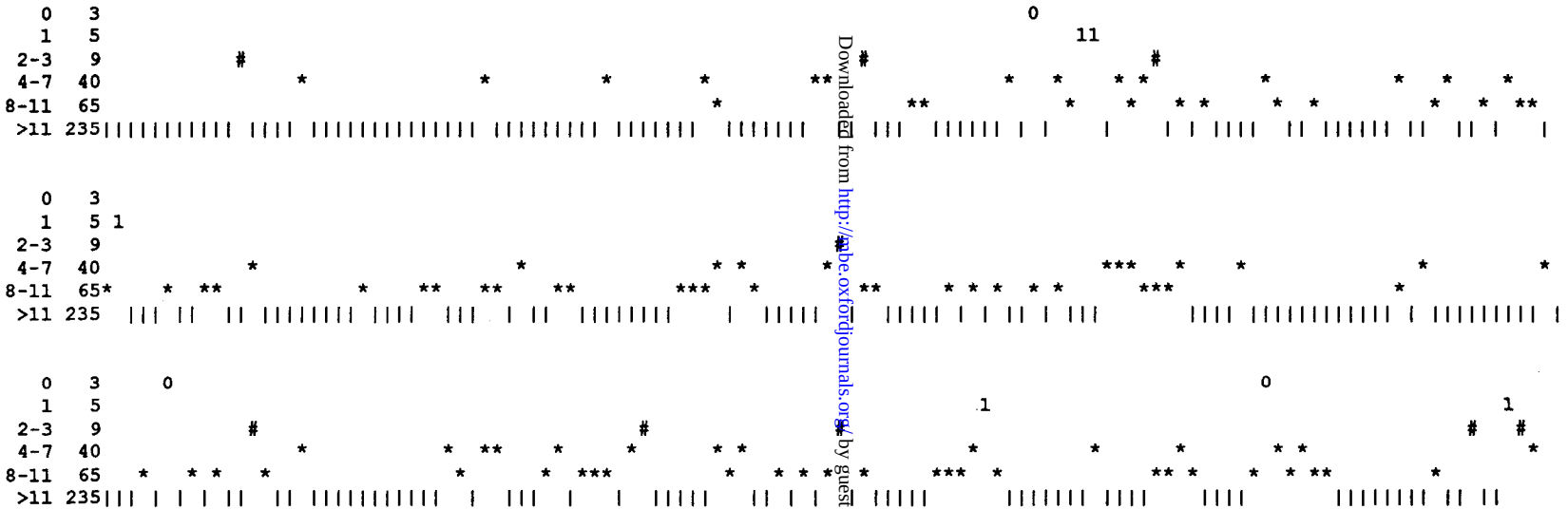

FIG. 5.—Limited variance of specific third bases. A set of 39 actin genes were aligned and the majority third base was determined for each codon. Excluded were methionine- and tryptophane-coding positions and all positions at which more than three amino acid replacements had occurred in the set. For the remaining positions the total number of third bases that differed from the majority was calculated, in order to classify the positions as more or less variable. Most positions had a moderately large number of variants, as expected for the distant set of genes shown in table 3. In this figure the coding region has been broken into three segments: codons 1–119, 120–239, and 240–357. On the bottom line of each segment are shown the majority of positions that have large variation. Above are lines showing five classes of less variable positions. The three positions in the top line do not vary at all and are marked "0." The next line shows five positions at which only one variant has been observed, marked "1." Below is a line showing nine positions that had two or three variants, marked "#." The next two lines show 40 and 65 positions with 4–7 and 8–11 variants, respectively, marked "*." The table repeated at the left shows, in the first column, the numbers of variants defining each class and, in the second column, the number of members of the class, for the whole gene. There are several distinct clusters of less variable positions.

strictions. Three of four gene sets examined include positions at which variation is rare or forbidden.

Implications

The implications of studies of compositional effects on synonymous similarity apply to several areas of research: use of synonymous substitutions for determination of the evolutionary relationships among genes and the species that are their hosts; attempts to determine the underlying mutation rates free of selection; and questions about the origins of codon bias. The implications are considered in this order. It has been pointed out elsewhere that the intergenic differences calculated for synonymous substitutions are affected when the base compositions of compared genes differ (Sharp 1991) or are not "stationary" (Saccone et al. 1990). The relationships shown in equation (1) and the examples in tables 1 and 2 show how striking this effect may be. Apparent relationships of genes may not be the result of recent common ancestors but instead may be due to sharing high GC3%. Even screening for related genes for conserved proteins by hybridization would be very likely to give a positive signal at great distance in cases of similar high or low GC3%, while it might fail at close distances if the GC3% differed. In extreme cases, the compositional difference may be more important than the evolutionary divergence of the genes.

In assessing the relationships among species, it is desirable to make interspecies gene sequence comparisons on the same basis, regardless of differences in base composition (or to correct for it). Furthermore, many interesting evolutionary questions, such as the branching during the mammalian radiation, involve distances so large that synonymous-substitution differences are large. During its evolution, each gene undergoes many synonymous changes. While the sequence may continue to change, there is a limit to its degree of synonymous divergence away from ancestral sequences. This is because third bases may mutate to any synonymous possibility, including a return to the original, a process often referred to as "reversion." More distant genes show differences in their third bases, rising toward a limit of difference. To judge the significance of large synonymous divergence, it is important to determine the limiting value, which is expressed by equation (1) as a function of the G+C content of third bases. Modifications of this equation will probably be developed to correct the limit for small effects that depend on aspects other than the G+C composition of third bases.

As the limit is approached, estimates of the number of substitutions that have occurred rise to large values that are sensitive to the limiting value in use. Corrections for reversion are customary (Li et al. 1985), but it seems likely that greater confidence will be possible if the limiting degree of divergence can be estimated directly from a relationship such as equation (1) and if uncertainty can be calculated for it. The average deviation shown in figure 3 gives an overall view of the accuracy. Genes with extreme GC3% show not only reduced limiting divergence but also reduced rate of substitution, because of restrictions on possible substitutions. That process maintains compositional bias, for whatever reason, and slows divergence. Corrections for this effect on apparent evolutionary distance are called for, particularly when the genes compared from different species have different GC3%, but the analysis of this problem is put off for future work.

The forbidden or rare substitutions form a body of slowly changing positions that also effectively reduce the limiting value of divergence and require additional corrections as shown in figure 4, even though they are a minority of positions. Not

only is the limiting value important, but, when the codon usage is restricted by high or low G+C composition, there is a reduced set of alternative codons, and this directly affects the rate of drift by synonymous substitution. Better methods of determining relationships by synonymous-substitution divergence will be required to directly report both this effect and the limit of divergence, as well as their uncertainties.

The underlying mutation rate is of interest, since the rate of variation and change in the genome is centrally important to evolution. The issues of drift and the "neutrality" of mutations have been a source of major controversy (e.g., see Kimura 1983, pp. 34–54). It may be that the most important part of genomic variation is caused by rearrangement of the DNA regions involved in the regulation of transcription, since this can produce novel processes. Nevertheless, mutation by base substitution remains a significant area of work. It originally seemed that synonymous changes in coding regions would give a good estimate for the underlying mutation rate. However, from evidence such as that presented above it is clear that corrections must be made both for codon usage patterns and for specific restrictions on change at certain bases. It seems that the best estimates will come from comparing genes with similar and middle-range values of GC3%, although the evidence shows that this is a small subset of possible actin gene comparisons. The effect of very high or low GC3% could create large errors in assessing the drift rate and the underlying mutation rate. Thus genes that are highly expressed or that may have other usage restrictions should be avoided for this purpose.

With regard to the underlying reason for the differences in GC content for various genes, a major issue in the literature is whether the codon bias is maintained by a "mutation bias" or is due to selection on the coding regions. Sharp (1991) takes the position that there is selection on codon bias in bacteria, while Wolfe et al. (1989) take the position that in mammals the third positions drift freely and the GC composition or codon bias is set by unequal mutation rates that vary between locations in the genome (Shields 1990). Mouchiroud and Gautier (1990) take yet a different position in interpreting mammalian gene comparisons. Presumably, during evolution, genes have achieved particular solutions to their regulatory and messenger RNA structure requirements, and in some cases (table 2) these solutions include high GC3%. At the same time, most genes have found different solutions that do not include extreme G+C composition of third bases. The GC3% varies so widely among actins that the average difference in the set of 1,275 comparisons is 25%. For those that have ~40% identity of third bases, the average difference in third bases is ~30%. This observation has implications for gene and genomic evolution, and a related observation among *Drosophila* genes (Shields et al. 1988) has been interpreted as evidence for selection among synonymous codons. However indirect are the mechanisms that set codon bias and third-base composition, a reasonable view is that selection based on gene function is important. The general requirements for regulation of transcription of the gene and for structure of messenger RNA do indicate that the codon bias and composition of the third bases are determined by selection, since the third bases appear to have a significant role. The evidence reported in this paper does show that selection acts to prevent substitutions at a minority set of specific positions in coding regions.

Acknowledgments

I wish to thank Eric H. Davidson, Temple F. Smith, and Tim Hunkapiller. The work was partially supported by ONR and NIH grants.

LITERATURE CITED

BAINS, W. 1987. Codon distribution in vertebrate genes may be used to predict gene length. J. Mol. Biol. **197**:379–388.

BERNARDI, G., and G. BERNARDI. 1985. Codon usage and genome composition. J. Mol. Evol. **22**:363–365.

BOYES, J., and A. BIRD. 1992. Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. EMBO J. **11**:327–333.

BRITTEN, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science **231**:1393–1398.

BULMER, M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics **129**:897–907.

BULMER, M., K. H. WOLFE, and P. M. SHARP. 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA **88**:5974–5978.

COOPER, D. N., and M. KRAWCZAK. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. **83**:181–188.

COULENDRE, C., J. H. MILLER, P. J. FARABAUGH, and W. GILBERT. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature **274**:775–780.

GRANTHAM, R., C. GAUTIER, M. GOUY, R. MERCIER, and A. PAVE. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8**:r49–r62.

GRUSKIN, K. D., T. F. SMITH, and M. GOODMAN. 1987. Possible origin of a calmodulin gene that lacks intervening sequences. Proc. Natl. Acad. Sci. USA **84**:1605–1608.

HIGGINS, D. G., and P. M. SHARP. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73**:237–244.

IKEMURA, T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. **158**:573–597.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

LAWRENCE, J. G., D. L. HARTL, and H. OCHMAN. 1991. Molecular considerations in the evolution of bacterial genes. J. Mol. Evol. **33**:241–250.

LI, W.-H., and D. GRAUR. 1991. Fundamentals of molecular evolution. Sinauer, Sunderland, Mass.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

LONG, M., and J. H. GILLESPIE. 1991. Codon usage divergence of homologous vertebrate genes and codon usage clock. J. Mol. Evol. **32**:6–15.

MOUCHIROUD, D., and C. GAUTIER. 1990. Codon usage changes and sequence dissimilarity between human and rat. J. Mol. Evol. **31**:81–91.

PEARSON, W. R., and D. J. LIPMAN. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85**:2444–2448.

SACCONE, C., C. LANAVE, G. PESOLE, and G. PREPARATA. 1990. Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol. **183**:570–583.

SHARP, P. M. 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium:* codon usage, map position, and concerted evolution. J. Mol. Evol. **33**:23–33.

SHARP, P. M., and W.-H. LI. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4**:222–230.

SHARP, P. M., D. C. SHIELDS, K. H. WOLFE, and W.-H. LI. 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. Science **246**:808–810.

SHIELDS, D. C. 1990. Switches in species-specific codon preferences: the influence of mutation biases. J. Mol. Evol. **31**:71–80.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS, and F. WRIGHT. 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5**: 704–716.

WOLFE, K. H., P. M. SHARP, and W.-H. LI. 1989. Mutation rates differ among regions of the mammalian genome. Nature **337**:283–285.

MASATOSHI NEI, reviewing editor

Received June 22, 1992; revision received July 27, 1992

Accepted July 27, 1992